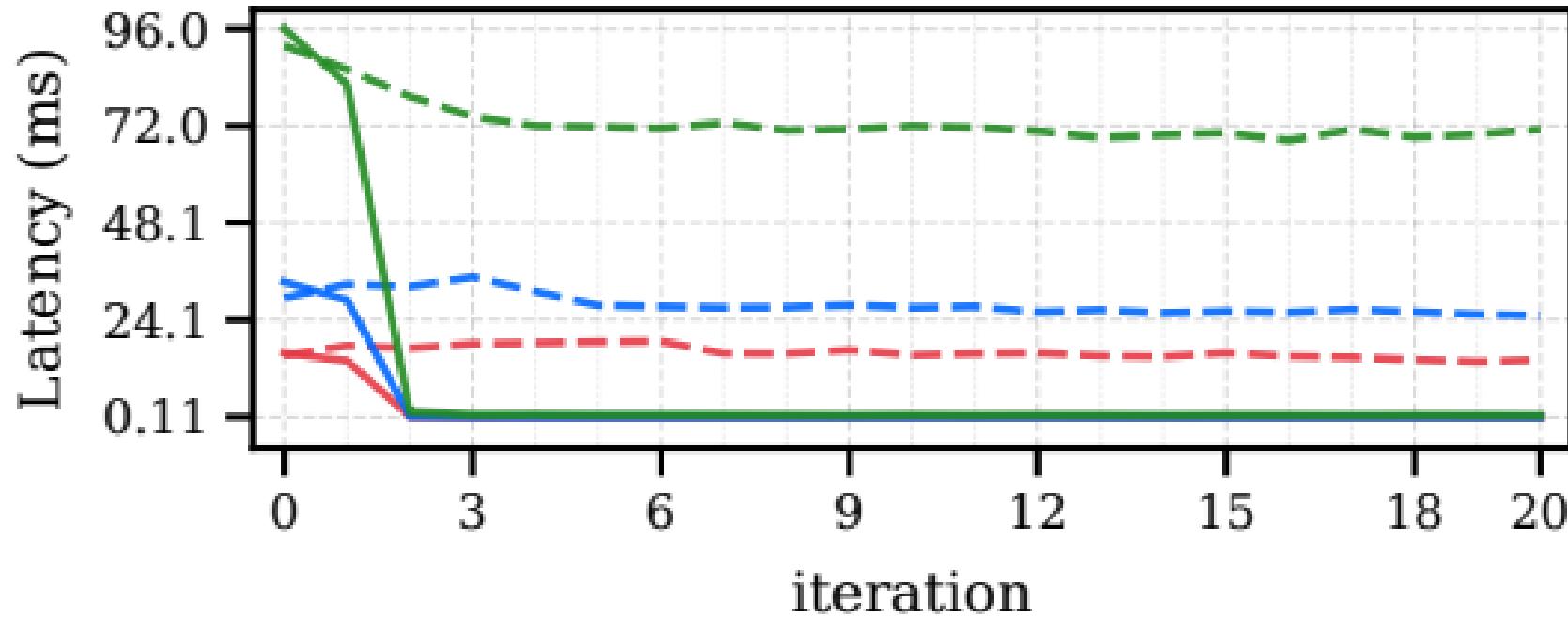


# Latency (P50)



Legend:

- Backend only (k=1)
- Backend only (k=10)
- With QVCache (k=1)
- With QVCache (k=100)